Data Science Interview Questions

PREDICTIVE MODELING

Q1. (Given a Dataset) Analyze this dataset and give me a model that can predict this response variable.

Q2. What could be some issues if the distribution of the test data is significantly different than the distribution of the training data?

Q3. What are some ways I can make my model more robust to outliers?

Q4. What are some differences you would expect in a model that minimizes squared error, versus a model that minimizes absolute error? In which cases would each error metric be appropriate?

Q5. What error metric would you use to evaluate how good a binary classifier is? What if the classes are imbalanced? What if there are more than 2 groups?

Q6. What are various ways to predict a binary response variable? Can you compare two of them and tell me when one would be more appropriate? What's the difference between these? (SVM, Logistic Regression, Naive Bayes, Decision Tree, etc.)

Q7. What is regularization and where might it be helpful? What is an example of using regularization in a model?

Q8. Why might it be preferable to include fewer predictors over many?

Q9. Given training data on tweets and their retweets, how would you predict the number ofretweets of a given tweet after 7 days after only observing 2 days worth of data?

Q10. How could you collect and analyze data to use social media to predict the weather?

Q11. How would you construct a feed to show relevant content for a site that involves userinteractions with items?

Q12. How would you design the people you may know feature on LinkedIn or Facebook?

Q13. How would you predict who someone may want to send a Snapchat or Gmail to?

Q14. How would you suggest to a franchise where to open a new store?

Q15. In a search engine, given partial data on what the user has typed, how would you predict the user's eventual search query?

greenstechnologys @ Porur

Q16. Given a database of all previous alumni donations to your university, how would you predict which recent alumni are most likely to donate?

Q17. You’re Uber and you want to design a heatmap to recommend to drivers where to wait for a passenger. How would you approach this?

Q18. How would you build a model to predict a March Madness bracket?

Q19. You want to run a regression to predict the probability of a flight delay, but there are flights with delays of up to 12 hours that are really messing up your model. How can you address this?

PROBABILITY

Q21. Bobo the amoeba has a 25%, 25%, and 50% chance of producing 0, 1, or 2 offspring, respectively. Each of Bobo’s descendants also have the same probabilities. What is the probability that Bobo’s lineage dies out?

Q22. In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the proba- bility that you see at least one shooting star in the period of an hour?

Q24. How can you get a fair coin toss if someone hands you a coin that is weighted to come up heads more often than tails?

Q25. You have an 50-50 mixture of two normal distributions with the same standard deviation. How far apart do the means need to be in order for this distribution to be bimodal?

Q26. Given draws from a normal distribution with known parameters, how can you simulate draws from a uniform distribution?

Q27. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

Q28. You have a group of couples that decide to have children until they have their first girl, afterwhich they stop having children. What is the expected gender ratio of the children that are born?What is the expected number of children each couple will have?

Q29. How many ways can you split 12 people into 3 teams of 4?

Q30. Your hash function assigns each object to a number between 1:10, each with equal probability. With 10 objects, what is the probability of a hash collision? What is the expected number of hash collisions? What is the expected number of hashes that are unused?

Q31. You call 2 UberX’s and 3 Lyfts. If the time that each takes to reach you is IID, what is theprobability that all the Lyfts arrive first? What is the probability that all the UberX’s arrive first?

greenstechnologys @ Porur

Q32. I write a program should print out all the numbers from 1 to 300, but prints out Fizz instead if the number is divisible by 3, Buzz instead if the number is divisible by 5, and FizzBuzz if the number is divisible by 3 and 5. What is the total number of numbers that is either Fizzed, Buzzed, or FizzBuzzed?

Q33. On a dating site, users can select 5 out of 24 adjectives to describe themselves. A match isdeclared between two users if they match on at least 4 adjectives. If Alice and Bob randomly pick adjectives, what is the probability that they form a match?

Q34. A lazy high school senior types up application and envelopes to n different colleges, but puts the applications randomly into the envelopes. What is the expected number of applications that went to the right college?

Q35. Let's say you have a very tall father. On average, what would you expect the height of his son to be? Taller, equal, or shorter? What if you had a very short father?

Q36. What's the expected number of coin flips until you get two heads in a row?

Q37. Let's say we play a game where I keep flipping a coin until I get heads. If the first time I get heads is on the nth coin, then I pay you 2n-1 dollars. How much would you pay me to play this game?

Q38. You have two coins, one of which is fair and comes up heads with a probability 1/2, and the other which is biased and comes up heads with probability 3/4. You randomly pick coin and flip it twice, and get heads both times. What is the probability that you picked the fair coin?

Data Analysis

Q39. Let's say you're building the recommended music engine at Spotify to recommend peoplemusic based on past lis- tening history. How would you approach this problem?

Q40. What is R2? What are some other metrics that could be better than R2 and why?

Q41. What is the curse of dimensionality?

Q42. Is more data always better?

Q43. What are advantages of plotting your data before performing analysis?

Q44. How can you make sure that you don't analyze something that ends up meaningless?

Q45. What is the role of trial and error in data analysis? What is the role of making a hypothesisbefore diving in?

Q46. How can you determine which features are the most important in your model?

Q47. How do you deal with some of your predictors being missing?

greenstechnologys @ Porur

Q48. You have several variables that are positively correlated with your response, and you thinkcombining all of the variables could give you a good prediction of your response. However, you see that in the multiple linear regression, one of the weights on the predictors is negative. What could be the issue?

Q49. Let’s say you’re given an unfeasible amount of predictors in a predictive modeling task. What are some ways to make the prediction more feasible?

Q50. Now you have a feasible amount of predictors, but you’re fairly sure that you don’t need all of them. How would you perform feature selection on the dataset?

Q51. Your linear regression didn’t run and communicates that there are an infinite number of best estimates for the regression coefficients. What could be wrong?

Q52. You run your regression on different subsets of your data, and that in each subset, the betavalue for a certain variable varies wildly. What could be the issue here?

Q53. Given that you have wifi data in your office, how would you determine which rooms and areasare underutilized and overutilized?

Q54. How would you quantify the influence of a Twitter user?

Q56. You have 5000 people that rank 10 sushis in terms of salt- iness. How would you aggregate this data to estimate the true saltiness rank in each sushi?

Q57. Given data on congressional bills and which congressio- nal representatives co-sponsored the bills, how would you determine which other representatives are most similar to yours in voting behavior? How would you evaluate who is the most liberal? Most republican? Most bipartisan?

Q58. How would you come up with an algorithm to detect plagiarism in online content?

Q59. You have data on all purchases of customers at a grocery store. Describe to me how you would program an algorithm that would cluster the customers into groups. How would you determine the appropriate number of clusters include?

Q60. In an A/B test, how can you check if assignment to the various buckets was truly random?

Q61. What might be the benefits of running an A/A test, where you have two buckets who areexposed to the exact same product?

Q62. What would be the hazards of letting users sneak a peek at the other bucket in an A/B test?

Q63. What would be some issues if blogs decide to cover one of your experimental groups?

Q64. How would you conduct an A/B test on an opt-in feature?

Q65. How would you run an A/B test for many variants, say 20 or more?

greenstechnologys @ Porur

Q66. How would you run an A/B test if the observations are extremely right-skewed?

Q67. I have two different experiments that both change the sign-up button to my website. I want to test them at the same time. What kinds of things should I keep in mind?

Q68. What is a p-value? What is the difference between type-1 and type-2 error?

Q69. How would you design an experiment to determine the impact of latency on userengagement?

Q70. What is maximum likelihood estimation? Could there be any case where it doesn't exist?

Q71. What's the difference between a MAP, MOM, MLE estimator? In which cases would you want to use each?

Q72. What is a confidence interval and how do you interpret it?

Q73. What is unbiasedness as a property of an estimator? Is this always a desirable property when performing inference? What about in data analysis or predictive modeling?

Q74. What is the difference between population and sample in data?

Q75. What is the difference sample and sample frame?

Q76. What is the difference between univariate, bivariate and multivariate analysis?

Q77. What is difference between interpolation and extrapolation?

Q78. What is precision and recall?

Q79. What is confusion matrix?

Q80. What is hypothesis testing?

Q81. What is a p-value in statistics?

Q82. What is difference between Type-I error and Type-II error in hypothesis testing?

Q83. QWhat are the different types of missing value treatment?

Q84. What is gradient descent?

Q85. What is difference between supervised and unsupervised learning algorithms?

Q86. What is the need for regularization in model building?

Q87. Difference between bias and variance tradeoff?

Q88. How to solve overfitting?

Q89. How will you detect the presence of overfitting?

Q90. How do you determine the number of clusters in k-means clustering?

greenstechnologys @ Porur

Q91. What is the difference between causality and correlation?

Q92. Explain normal distribution?

Q93. What are the different ways of performing aggregation in python using pandas?

Q94. What are merge two list and get only unique values?

Q95. How to save and retrieve model objects in python?

Q97. What is an ensemble learning?

Q98. Name few libraries that is used in python for data analysis?

Q99. What are the different types of data?

Q100. What is a lambda function in python?

Q101. What are the different sampling methods?

Q102. Common Data Quality Issues

Q103. What is the difference between supervised learning and un-supervised learning?

Q104. What is Imbalanced Data Set and how to handle them? Name Few Examples?

Q105. If you are dealing with 10M Data, then will you go for Machine learning (or) Deep learning Algorithm?

Q106. Examples of Supervised learning algorithm?

Q107. In Logistic Regression, if you want to know the best features in your dataset then what you would do?

Q108. What is Feature Engineering? Explain with Example?

Q109. How to select the important features in the given data set?

Q111. What is Variance inflation Factors (VIF)

Q112. Examples of Parametric machine learning algorithm and non-parametric machine learning algorithm

Q113. What are parametric and non-parametric machine learning algorithm? And their importance

Q114. When does linear and logistic regression performs better, generally?

Q115. Why you call naïve bayes as “naïve” ?

Q116. Give some example for false positive, false negative, true positive, true negative

Q117. What is Sensitivity and Specificity?

greenstechnologys @ Porur

greenstechnologys @ Porur

Q118. When to use Logistic Regression and when to use Linear Regression?

Q119. What are the different imputation algorithm available?

Q120. What is AIC(Akaike Information Criteria)

Q121. Suppose you have 10 samples, where 8 are positive and 2 are negative, how to calculate Entropy (important to know)

Q122. What is perceptron in Machine Leaning?

Q123. How to ensure we are not over fitting the model?

How the root node is predicted in Decision Tree Algorithm?

Q125. What are the different Backend Process available in Keras?

Q126. Name Few Deep Learning Algorithm

Q127. How to split the data with equal set of classes in both training and testing data?

Q128. What do you mean by giving “epoch = 1” in neural network?

Q129. What do you mean by Ensemble Model? When to use?

Q130. When will you use SVM and when to use Random Forest?

Q131. Applications of Machine Learning?

Q132. If you are given with a use case – ‘Predict whether the transaction is fraud (or) not fraud”, which algorithm would you choose

Q133. If you are given with a use case – ‘Predict the house price range in the coming years”, which algorithm would you choose

Q134. What is the underlying mathematical knowledge behind Naïve Bayes?

Q135. When to use Random Forest and when to Use XGBoost?

Q136. If you are training model gives 90% accuracy and test model gives 60% accuracy? Then what problem you are facing with?

Q137. In Google if you type “How are “it gives you the recommendation as “How are you “/”How do you do”, this is based on what?

Q138. What is margin, kernels, Regularization in SVM?

Q139. What is Boosting? Explain how Boosting works?

Q140. What is Null Deviance and Residual Deviance (Logistic Regression Concept?)

Q141. What are the different method to split the tree in decision tree?

Q142. What is the weakness for Decision Tree Algorithm?

Q143. Why do we use PCA(Principal Components Analysis) ?

Q144. During Imbalanced Data Set, will you

Q145.How to ensure we are not over fitting the model?

Q146. Steps involved in Decision Tree and finding the root node for the tree

Q147. What is hyper plane in SVM?

Q148. Explain Bigram with an Example?

Q149. What are the different activation functions in neural network?

Q150. Which Algorithm Suits for Text Classification Problem?

Q151. You are given a train data set having lot of columns and rows. How do you reduce the dimension of this data?

Q152. You are given a data set on fraud detection. Classification model achieved accuracy of 95%.Is it good?

Q153. What is prior probability and likelihood?

Q154. How can we know if your data is suffering from low bias and high variance?

Q155. How is kNN different from kmeans clustering?

Q156. Random Forest has 1000 trees, Training error: 0.0 and validation error is 20.00.What is the issue here?

Q157. Data set consisting of variables having more than 30% missing values? How will you deal with them?

Q158. What do you understand by Type I vs. Type II error?

Q159. Based on the dataset, how will you know which algorithm to apply ?

Q160. Why normalization is important?

Q161. What is Data Science?

Q162. What is Machine Learning?

Q163. What is Deep Learning?

Q164. Where to use R & Python?

Q165. Which Algorithms are used to do a Binary classification?

Q166. Which Algorithms are used to do a Multinomial classification?

greenstechnologys @ Porur

Q167. What is LOGIT function?

Q168. What are all the pre-processing steps that are highly recommended?

Q169. What is Normal Distribution?

Q170. What is empirical Rule?

Q171. What is Regression problem statement?

Q172. What are all the Error metrics for Regression problem statement?

Q173. What is R value in Linear regression?

Q174. What is an Outlier?

Q175. What are all the mechanisms which can identify Outliers?

Q176. How can we treat Outliers?

Q177. What are all the standard imputations that can be carried for missing value treatments?

Q178. What is the formula for calculating Upper whisker & Lower whisker value in Box plot?

Q179. What is the skewed Distribution & uniform distribution?

Q180. What is the key assumption for Naive Bayes?

greenstechnologys @ Porur